# Compositional generalization through abstract representations in human and artificial neural networks

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Humans have a remarkable ability to rapidly generalize to new tasks that is difficult to reproduce in artificial learning systems. Compositionality has been proposed as a key mechanism supporting generalization in humans, but evidence of its neural implementation and impact on behavior is still scarce. Here we study the computational properties associated with compositional generalization in both humans and artificial neural networks (ANNs) on a highly compositional task. First, we identified behavioral signatures of compositional generalization in humans, along with their neural correlates using whole-cortex functional magnetic resonance imaging (fMRI) data. Next, we designed pretraining paradigms aided by a procedure we term *primitives pretraining* to endow compositional task elements into ANNs. We found that ANNs with this prior knowledge had greater correspondence with human behavior and neural compositional signatures. Importantly, primitives pretraining induced abstract internal representations, excellent zero-shot generalization, and sample-efficient learning. Moreover, it gave rise to a hierarchy of abstract representations that matched human fMRI data, where sensory rule abstractions emerged in early sensory areas, and motor rule abstractions emerged in later motor areas. Our findings give empirical support to the role of compositional generalization in human behavior, implicate abstract representations as its neural implementation, and illustrate that these representations can be embedded into ANNs by designing simple and efficient pretraining procedures.

## 1 Introduction

Humans can efficiently transfer prior knowledge to novel contexts, an ability commonly referred to as transfer learning. One proposed mechanism underlying transfer learning is compositional generalization (or compositional transfer) – the ability to systematically recompose learned concepts into novel concepts (e.g., "red" and "apple" can be combined to form the concept of a "red apple") [5, 8, 17]. Indeed, it has been suggested that an algorithmic implementation of compositional generalization is one of the key missing ingredients that ANN models need in order to achieve human-like learning and reasoning capabilities [27, 25]. Therefore, quantifying how compositional generalization is manifested in human behavior and investigating its underlying implementation in biological brains is a natural first step to harness and deploy it in machine learning models.

Recent studies that investigated compositionality in machine learning have typically relied on architectures comprised of specialized modules. For instance, disentangled representation learning separates

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

the independent factors underlying the structure of the input data into disjoint components of the feature vector [14, 15, 32, 13]. Program synthesis methods achieve state-of-the-art performance on systematic generalization [17] through model architectures built by combining specialized neural and symbolic program modules interacting to search over a space of valid production rules [26, 34, 41].

Complementing these studies, *abstract representations* have been recently proposed as vector representations that reconcile compositional generalization with distributed neural codes [2]. In particular, *parallel abstract representations* – representations with a high Parallelism Score as previously defined [2] – support out-of-context generalization by encoding changes in individual variables as a linear shift in the representations. This notion of abstraction implies that these representations are compositionally additive; novel compositions are encoded as the vector sum of distinct abstract representations. This is similar to how word2vec embeddings solve relational analogy tasks [31, 28] and generalizes disentangled representations by allowing for arbitrary affine transformations of disentangled codes. Crucially, this type of representation is operationally defined in a way that can be quantified in neuroimaging data by computing the Parallelism Score metric defined in [2]. In other words, parallel abstract representations are a computationally promising candidate as neural substrate implementing compositional generalization, and are also measurable in the human brain by computing the Parallelism Score across fMRI voxels during neuroimaging experiments.

This work is motivated by the working hypothesis that parallel abstract representations support compositional generalization. Accordingly, we first characterized the behavioral signatures of compositional generalization in a task that systematically varied rule conditions across 64 contexts, showing that humans generalize better to tasks with greater similarity structure to previous tasks. We then analyze fMRI imaging data showing that parallel abstract representations are distributed across the entire cortex in a content-specific way during the execution of our compositional task. This supports our working hypothesis that parallel abstract representations may implement compositional generalization. We then design a pretraining paradigm for ANNs to emulate humans' prior knowledge about the compositional task elements, finding that ANNs pretrained in this way exhibit 1) more abstract representations, 2) excellent generalization performance, and 3) sample-efficient learning. Finally, we find that the layerwise organization of abstract representations in pretrained ANNs recapitulates the content-specific distribution in human cortex. Together, these findings provide empirical evidence for the role of abstract representations in supporting compositional generalization.

### 1.1 Related work

Several recent studies in neuroscience have applied analytic tools to identify the neural basis of rapid generalization in biological neural networks. Such studies employed various measures – cross-condition generalization [2, 36, 7, 4], state-space projections of task-related compositional codes [46, 38, 22], and Parallelism Score [2] – to quantify the generalizability and abstraction of representations. Prior work in neuroscience has primarily evaluated compositionality in limited context settings (e.g., up to 10 contexts), or without manipulating different types of features (e.g., higher-order vs. sensory/motor features). Moreover, these neuroscience studies used simple task paradigms due to limitations in either the model organism (rodents and monkeys are unable to perform complex tasks [2]) or to isolate specific types of abstraction in humans (e.g., logical abstractions [36]). Here we significantly expand on prior work by using a 64-context compositional task that systematically varies different types of task features (e.g., sensory, motor, and logical rules) to evaluate content-specific abstractions across the entire brain and multilayer ANNs. This work also complements related work in compositional generalization in machine learning [25, 41, 17, 26, 40, 45, 16]. However, those studies primarily focused on building models that improve on current compositional generalization benchmarks on arbitrarily complex compositional tasks, such as SCAN [25], COG [45], or GQA [16]. Importantly, these studies did not directly benchmark ANN behavior (or representations) against human behavioral and neural data, making a direct comparison difficult. Here we leveraged a non-trivial 64-context compositional paradigm to investigate the representational principles that facilitate compositional generalization in both humans and ANNs.

## 2 Methods

### 2.1 The Concrete Permuted Rule Operations (C-PRO) task paradigm

We used the C-PRO paradigm (Fig. 1a) during fMRI acquisition and ANN model training. Briefly, the C-PRO paradigm permutes specific task rules from three rule domains (logical decision, sensory semantic, and motor response) to generate dozens of novel task contexts. This creates a context-rich dataset in the task context domain. The sensory rule indicates which stimulus feature should be attended to. The logic rule specifies a Boolean operation to be implemented on the stimulus feature set. The motor rule specifies a specific motor action (i.e., a button press with a specific finger). One of 256 possible unique stimulus combinations could be presented with each task context. Visual dimensions included either horizontal or vertical bars with either blue or red coloring. Auditory dimensions included continuous (constant) or beeping high or low pitched tones.

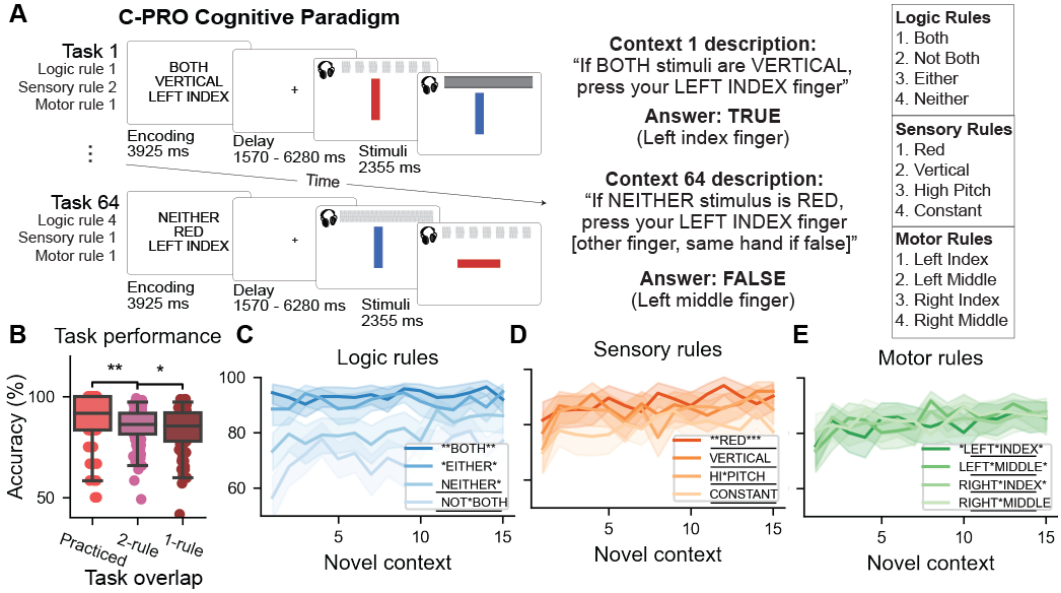


Figure 1: a) The C-PRO paradigm permutes 12 rules belonging to three different rule domains – logical, sensory, and motor gating – to generate up to 64 unique contexts. b) Human performance on novel task contexts was significantly lower than the practiced contexts (participants were trained on four practice contexts prior to the test session). Moreover, subjects performed novel task contexts with more rule overlap with practiced contexts at a higher accuracy. c-e) Task performance as a function of task trials for each rule (novel contexts only). Consistent with compositional generalization, participants had a significant increase in task performance in 10/12 rules, even though each rule was used in a novel context. Shaded area around line plots (c-e) reflects the 95% confidence interval.

Each rule domain (logic, sensory, and motor) consists of 4 specific rules (Fig. 1a). A task context is comprised of one rule from each domain, for a total of 64 possible task contexts (4 logic x 4 sensory x 4 motor). Subjects were trained on 4/64 "practiced" contexts prior to the fMRI session. The 4 practiced contexts were selected such that all 12 rules were equally practiced. Subjects' mean performance across all trials was 84% (median=86%; chance=25%). See Appendix A.2 for details.

### 2.2 The geometry of abstract neural representations

Behavioral signatures of compositional generalization can be investigated by measuring behavioral performance as a function of task composition and prior learning. Neural signatures of generalization can be identified using analysis methods that characterize the geometry of neural activations during task generalization. In particular, prior work proposed the Parallelism Score (PS) [2] as a measure to evaluate the consistency of task variable representation across contexts. Intuitively, PS identifies a coding axis (i.e., the parallel displacement vector) across task contexts that aids generalization.

We posit that representations with high PS (the specific type of abstract representation we investigate) support compositional generalization in human behavior. We illustrate here how PS is reflected in the geometry of neural representations with respect to the rule domains of the C-PRO task. Let us consider a set of C-PRO contexts with logic rules BOTH or EITHER, and sensory rules with values RED or VERTICAL (Fig. 2). High PS in the logic rule domain indicates that the difference in activation vectors between contexts with BOTH and EITHER rules is the same when paired with either the RED or VERTICAL sensory rules. Thus, a change from BOTH to EITHER results in the same parallel change irrespective of the sensory rule (Fig. 2c). In contrast to unstructured high-dimensional representations (Fig. 2a), this would afford high generalization, since the effect of changing the logic rule in either sensory rules automatically transfers to the other sensory rule.

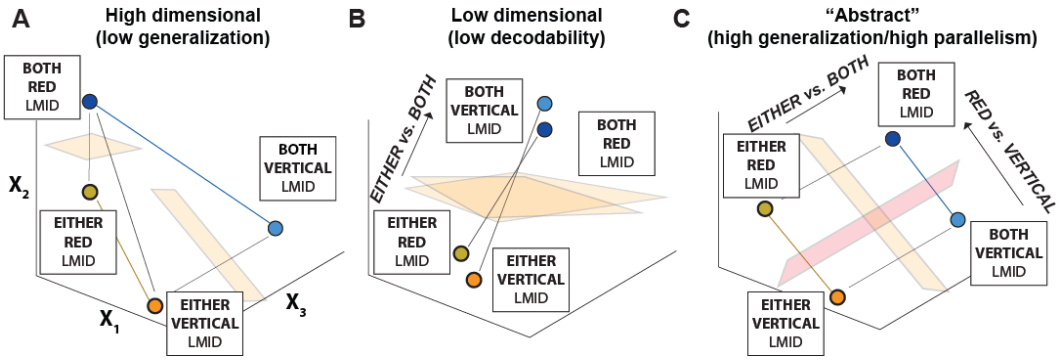


Figure 2: Hypothetical geometric configurations of neural activation space for "BOTH vs. EITHER" and "RED vs. VERTICAL" rule contrasts. a) High-dimensional representations of task activations lead to low PS (in addition to low generalizability across conditions) of rules. b) Low-dimensional representations lead to overall low decodability, but some generalizability (across limited features). c) Parallel Abstract Representation of the neural activations leads to high generalizability.

### 2.3 Parallelism score

We generalize the definition of PS by [2] to tasks where variables can assume an arbitrary number of values (as opposed to being binary) and applied it to human fMRI and internal ANN activations. PS is defined as the cosine angle of the coding directions of the same rules in different contexts in the neural activation space (e.g., voxels or neurons within a brain region) (Fig. 3a-c). A cosine angle close to 1 indicates coding directions that are highly parallel, despite differences in context. We compute the coding angle for a specific rule dichotomy (e.g., the coding direction "BOTH" vs. "EITHER") by identifying all pairs of task contexts that had exactly the same secondary (sensory) and tertiary (motor) rules. For each pair, we subtracted the fMRI voxel activation vectors associated with each context to obtain the vector that represented that coding direction (see Fig. 3a). We did this for all other pairs in that coding direction. Defining $v_i$ as this coding vector for the $i$th pair, we computed the PS score for one dichotomy as $PS_k = \frac{1}{16}\sum_{i \neq j}^{16} cos(v_i, v_j))$, since there are 16 possible pairs for each coding direction within the C-PRO task. To obtain the PS for a specific rule domain (e.g., logic, sensory, or motor rules), $PS_k$ is computed for every coding direction, then averaged (e.g., for logic PS, the average of "BOTH" versus "EITHER", "BOTH" versus "NEITHER", etc.).

Statistical testing was performed using a non-parametric procedure, where we shuffled labels within each rule domain 1000 times and re-calculated PS to produce a null distribution. We corrected for multiple comparisons (across brain regions) using non-parametric family-wise error correction [33].

### 2.4 ANN construction and training

The primary ANN architecture had two hidden layers (128 units each) and an output layer that was comprised of four units that corresponded to each motor response (see Appendix section A.7 and Fig. 8 for additional details). Training used a cross-entropy loss function and the Adam optimizer [24]. The ANN transformed the trial input vector into a 4-element response vector with the

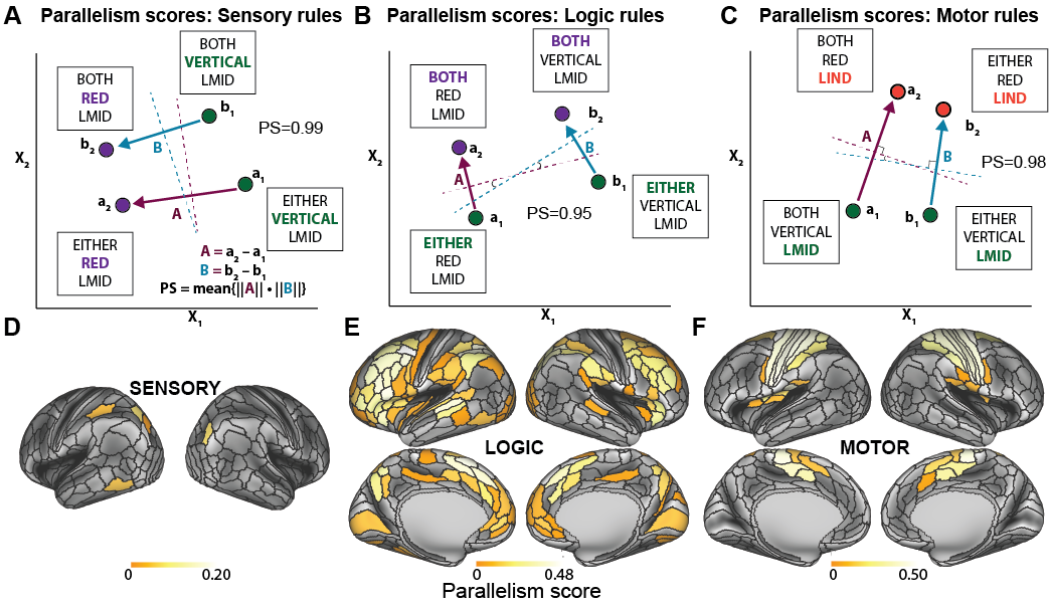


Figure 3: a-c) 2-D schematic visualization of PS estimation for the a) sensory, b) logic, and c) motor rule domains for a specific rule pair (e.g., RED vs. VERTICAL). Intuitively, PS captures the geometry of the neural activation space (ANN or fMRI data) by measuring the cosine angle between two linear decoders trained to distinguish two rule conditions in different task contexts. d-f) PS was calculated for each rule domain for every brain region [10]. PS was highest in association areas for logic rules, dorsal attention network regions for sensory rules, and somatomotor network for motor rules.

equation $Y = f_{ReLU}(X_h W_h + b_h)$. Weights and biases were initialized from a uniform distribution $U(-\sqrt{1/k}, \sqrt{1/k})$, where $k$ is the number of input features from the previous layer.

Training on the C-PRO task was performed in a sequential learning paradigm. To mimic the human experiment, an arbitrary set of four practiced contexts was initially selected for training. (This was randomly selected across different ANN initializations.) Then, novel task contexts were incrementally added into the set of training contexts.

# 3 Results

## 3.1 Behavioral signatures of rapid compositional generalization in humans

We evaluated human behavioral compositional generalization by assessing performance on novel contexts in the C-PRO paradigm. Since adult humans have decades of prior knowledge, subjects were able to compositionally generalize to novel task contexts without any training (novel accuracy=84.17%, chance=25%, Wilcoxon signed-rank p<0.0001). However, subjects performed the four practiced contexts better than novel contexts (practiced=87.67%, novel=84.17%; p=0.003). We next assessed how performance on novel contexts changed as a function of shared rule structure to the practiced contexts. Consistent with compositional transfer of previously learned rules, performance on novel task contexts improved as a function of similarity to the practiced contexts (accuracy, 2-rule overlap=84.86%; 1-rule overlap=83.48%; practiced vs. 2-rule overlap, p=0.008; 2-rule vs. 1-rule overlap, p=0.03; Fig. 1b). Though our findings are consistent with compositional transfer, we found that rapid transfer to novel contexts is more difficult. However, we found that increased exposure to specific rules improved performance on subsequent novel contexts using that same rule (all except for the "Both" and "Either" rules, likely due to ceiling effects, FDR-corrected p<0.05; Fig. 1c-e). This suggests that even though performance in novel contexts is worse than practiced contexts, subjects can improve rule transfer with increased practice (or pretraining).

### 3.2 Spatial and content-specific topography of abstract representations in human cortex

We extended prior work to identify abstract representations using PS across the entire human cortex [2, 4, 36]. We calculated PS for each rule domain separately (Fig. 3a-c) using the vertices/voxels within each parcel (i.e., brain region) as activation vectors. We found topographic differences of sensory, logic, and motor rule abstractions tiled across human cortex (Fig. 3d-f). Specifically, we found that statistically significant sensory rule abstractions were primarily identified in higher order visual areas and the dorsal attention network (i.e., brain areas involved in the top-down selection of visual stimuli) (PS of significant regions=0.15; family-wise error (FWE)-corrected $p<0.05$; Fig. 3d). Logic rule abstractions were more widely distributed, but primarily observed in frontoparietal areas (PS of significant regions=0.22; FWE-corrected $p<0.05$; Fig. 3e). Motor rule abstractions were primarily localized to somatomotor cortex (PS of significant regions=0.29; FWE-corrected $p<0.05$; Fig. 3f). Notably, regions with abstract representations form a subset of regions of those that contain rule information using standard decoding methods (Fig. 7). This ensures that high PS is accompanied by highly decodable representations (i.e., high dimensionality).

### 3.3 Embedding prior knowledge into ANNs with simple pretraining tasks

Human behavioral data suggested improved compositional generalization with increased task rule exposure, in addition to the years of "pretraining" from ordinary development (i.e., at least 18+ years). Thus, we sought to evaluate whether embedding prior knowledge of rules could improve compositional generalization in ANNs, while simultaneously investigating how prior knowledge impacts the geometry of ANNs' internal task representations. Given that the C-PRO task was specifically designed as a compositional task that conjoined three task rules, we created pretraining paradigms designed to teach ANNs basic rule knowledge (Fig. 4; see Appendix for full description).

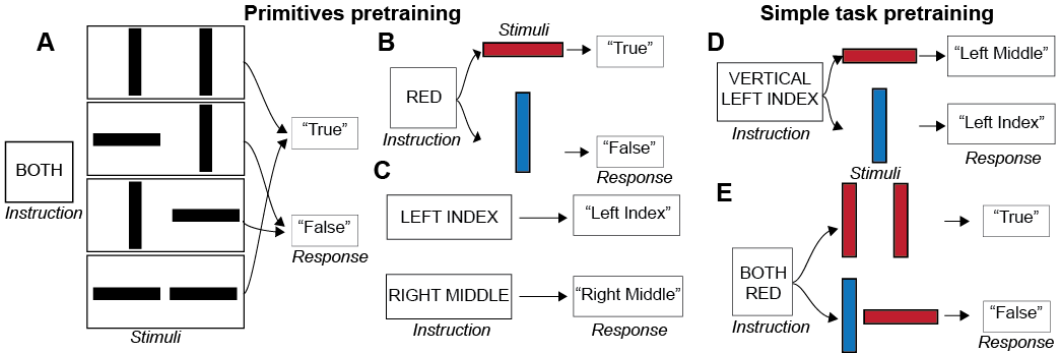


Figure 4: a) The logic rule primitives task involved teaching boolean relations among different logical operations. For example, when presented with the "BOTH" rule, the task was to distinguish two identical ("True") versus two different ("False") stimuli (i.e., same vs. different). b) Sensory rules involved mapping sensory rules onto stimulus features. c) Motor rules involved mapping motor rules onto motor output units. d-e) Simple task pretraining (2-rule tasks) was designed to teach the model how to perform simple (d) sensorimotor mappings and (e) logical-sensory gatings.

We constructed a simple feedforward ANN with two hidden layers (Appendix A.7; Fig. 8). This made it easier to investigate the effects of pretraining on internal representations, rather than architectural choices. We designed two pretraining paradigms: Primitives (1-rule) and Simple task (2-rule) pretraining. Primitives pretraining trained on 1-rule tasks that focused explicitly on learning the semantics of primitive rule features (Fig. 4a-c). This included distinguishing sensory stimuli, learning motor response mappings (e.g., "left index" rule would lead to a left index response), and abstract logical relations, which involved learning the boolean relations amongst logic rules. Simple task pretraining focused on learning 2-rule conjunctions (i.e., a sensory and motor rule pairing / logical and sensory rule pairing) (Fig. 4d-e). Importantly, these pretraining paradigms focused on learning primitive 1- or 2-rule associations that were significantly simpler than the full C-PRO task (3-rule combination).

### 3.4 Pretraining induces abstractions, zero-shot performance, and sample efficiency

We measured the PS in ANNs trained with different pretraining routines: Vanilla (no pretraining), Primitives pretrained, Simple task pretrained, and Combined (Primitives + Simple task pretrained). Each pretraining condition was trained until ANNs achieved 99% accuracy (see A.8 for details). PS was calculated for each rule domain using the ANN's hidden layer activations (see A.9). Pretrained ANNs had higher PS than the Vanilla ANN (Primitives vs. Vanilla, t(37)=5.26, p=1e-05; Simple task vs. Vanilla, t(37)=8.46, p=1e-11; Combined vs. Vanilla, t(37)=3.03, p=0.003) (Fig. 5a). Moreover, PS increased from Primitives to Simple task pretraining (t(37)=3.91, p=0.0002), though no significant increase in PS was observed in Combined vs. Simple task pretraining.

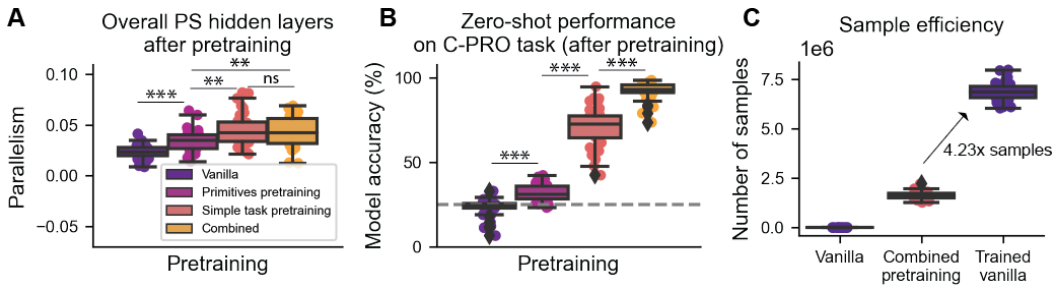


Figure 5: a) PS of hidden units averaged across all rule domains. b) Zero-shot learning of all 64 C-PRO contexts. c) Sample efficiency of models (Combined and trained vanilla model were performance-matched). Total samples, including pretraining samples (if applicable).

We next evaluated the zero-shot performance on the full C-PRO task after pretraining (Fig. 5b). As expected, the Vanilla ANN performed near chance (acc=23.25%, chance=25%, one-sided t(38)=-2.17, p=0.98). Primitives pretraining marginally improved zero-shot performance (acc=31.51%, t(38)=8.09, p<1e-9). Simple task pretraining exhibited significant improvement over Primitives pretrained models (acc=70.57%, Simple task vs. Primitives, t(37)=19.84, p<1e-31). Finally, we found that Combined pretraining had excellent zero-shot performance on the entire C-PRO task (acc=92.15%, Combined vs. Simple task pretraining, t(37)=10.85, p<1e-16). Notably, we found that PS and zero-shot performance monotonically increased with pretraining, illustrating that classic multilayer networks can transfer abstract representations for systematic zero-shot generalization [17].

Finally, we sought to assess the impact of pretraining on learning/sample efficiency. We therefore trained a Vanilla network (no pretraining) on 60/64 C-PRO contexts to match the zero-shot performance of the Combined pretraining model (i.e., at least 90% accuracy on the 60 context training set). We found that on the remaining test set (4/64 C-PRO contexts), the Vanilla trained model achieved 96.02% generalization performance, but required up to 4.23x training samples to match the performance of the Combined model (Fig. 5c). Critically, the 4.23x more training samples included all possible samples (pretraining and C-PRO samples). This illustrated that pretraining afforded both zero-shot generalization and sample efficient learning.

### 3.5 Pretraining leads to compositional generalization in ANNs comparable to humans

We evaluated the learning and generalization dynamics of ANNs with and without pretraining, after training ANNs on 4 of the full C-PRO contexts. This matched the human experiment, since humans were exposed to 4 "practice" contexts prior to performing the remaining 60 novel contexts. (ANN training was stopped after achieving 90% performance on the 4 practiced contexts.) We found overall poor generalization on novel task contexts in the Vanilla model (accuracy, practiced=94.37%, novel=28.79%; p<0.0001; Fig. 9a). This suggested that ANNs with no prior knowledge cannot compositionally generalize. We subsequently compared generalization performance on ANNs after pretraining. We found that with Primitives pretraining, generalization performance significantly improved (57.97%; Fig. 9b). We observed additional improvements with Simplified task pretraining (86.79%; Fig. 9c), achieving generalization performance on par with human performance (Fig. 9d).

We next incrementally trained all ANN models on novel contexts, by adding one novel context into the training set at a time. We tested generalization performance on the held-out (test set) contexts until ANNs were trained on 63/64 contexts (Fig. 10a). Generalization performance on novel contexts was significantly higher in ANNs with either pretraining routine (Fig. 10b). This was despite the fact that all ANNs had the same stopping criteria (i.e., 90% accuracy on the C-PRO training set). We ran an additional experiment where each of the ANNs were shown a fixed number of C-PRO task samples during training, replicating our core finding (Fig. 11). These findings suggest that the inductive biases formed during pretraining significantly improve downstream generalization performance.

### 3.6 Pretraining ANNs facilitates sample-efficient learning throughout novel task learning

We sought to evaluate how pretraining impacted sample efficiency. We found that pretrained ANNs became more sample efficient as the training set expanded, even when accounting for total number of (pretraining and C-PRO) samples (Fig. 10b). We quantified the generalization performance to sample efficiency ratio as the generalization inefficiency, finding that after learning only 7 C-PRO contexts, vanilla ANNs generalized worse than pretrained ANNs (Fig. 10c). These findings support the notion that pretraining can simultaneously improve compositional generalization and sample efficiency.

### 3.7 Convergent hierarchy of abstract representations in humans and ANNs

Analysis of human fMRI data revealed that content-specific abstraction was spatially heterogeneous across cortex. Recent neuroscience work has identified hierarchical gradients that organize along a sensory input-to-motor output axis in both resting-state [29] and multi-task fMRI data [18]. We therefore sought to quantify PS across the sensory-to-motor hierarchy in fMRI data, and compare it to PS changes in the feedforward hierarchy (i.e., layer-depth) in ANNs. We focused our analyses on the Combined pretrained model (which incorporates both Primitives and Simple task pretraining) due to its excellent zero-shot generalization (Fig. 5b). In addition, we extended our model to include three hidden layers to make it easier to compare PS of different hidden-layer depths to the three cortical systems of interest: sensory, association, and motor systems (Fig. 6a).

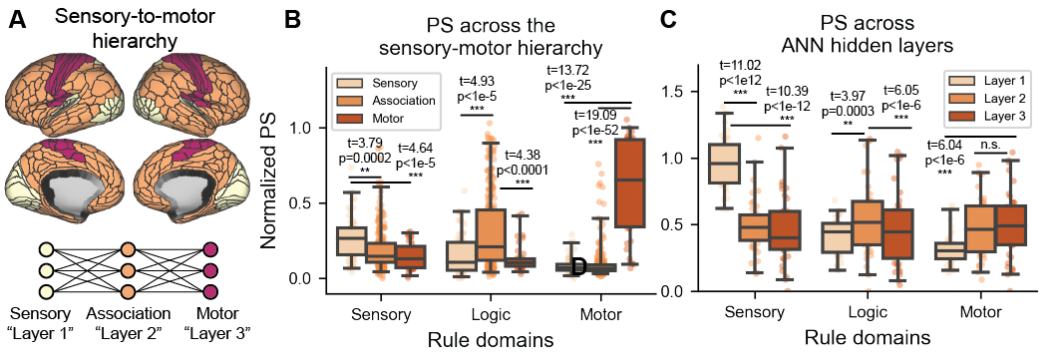


Figure 6: a) A discretized sensory-to-motor hierarchy (see Fig. 14 for discretization details). b) We computed the normalized PS (i.e., the PS of each brain region normalized by the maximum PS across all regions) for each rule domain across the discretized cortical systems. c) Same analysis as in b), but using the PS found in each ANN hidden layer.

We measured the PS for each rule domain for sensory, association, and motor systems. Sensory rule PS was highest in the sensory system, logic rule PS was highest in association systems, and motor rule PS was highest in the motor system (Fig. 6b). To observe whether similar hierarchical PS organization emerged in ANNs, we used the Combined pretrained model with three hidden layers, and plotted PS as a function of ANN depth. Since our ANN transformed sensory inputs into motor outputs, we analogized each ANN layer to the sensory, association, and motor cortical systems (Fig. 6a). We found a similar pattern in the ANN: sensory PS peaked in the first hidden layer; logic PS peaked in the second hidden layer; and motor PS peaked in the last two hidden layers

(Fig. 6d). We corroborated these findings using a continuous sensory-motor hierarchical gradient map (without discretization) (Fig. 13-14). (In addition, see Fig. 15-17 for PS scores for all possible rule dichotomies in human fMRI data and ANN activations.) These findings suggest that abstraction emerges as a function of rule-dependent specialization and hierarchical organization.

# 4 Discussion, Limitations, Conclusions

We provide empirical support for the role of compositionality in human generalization, and implicate abstract representations as its neural implementation. In classic ANNs, which are known to perform poorly during systematic generalization [17, 6], we found that computationally cheap pretraining paradigms embedded abstract representations that led to human-like generalization performance and sample efficient learning. When mapping abstract representations across cortex and ANN layers, we found converging patterns of rule-specific abstractions from early sensory areas/layers to late motor areas/layers across human and ANN hierarchies. These results reveal the hierarchical organization of content-specific abstractions in the human brain and ANNs, while revealing the impact of these abstractions for compositional generalization in models.

Our pretraining approach directly leverages knowledge of task structure to design pretraining routines that embed task biases into ANNs. Despite the sample efficiency of this approach, this pretraining approach requires the initial overhead of designing paradigms useful for downstream learning. A related approach that similarly requires prior knowledge of task structure is "representational backpropagation" – a regularization approach that aims to produce an idealized hidden representation [23]. However, there are other inductive bias approaches that do not require prior task knowledge. One approach constrains ANNs to produce abstract task representations by initializing ANN weights from a low-norm distribution [7]. However, initializing ANN weights in this regime is computationally costly. Another approach is to initialize networks with built-in modular structures to facilitate the re-use of network modules across tasks [30, 39]. However, exactly how such networks disentangle representations has not yet been explored. Nevertheless, all these approaches are complementary to each other. It will be important for future work to assess how these approaches may synergistically interact to optimize for sample-efficient generalization in multi-task settings.

Though we provide comprehensive evidence of the role of abstraction in compositional generalization, there are several limitations in the present study that future research can explore. We found that the spatial topography of abstract representations was highly content-dependent. However, analyses were limited to cross-context manipulations of limited rule types (sensory, logic, and motor gating), without addressing the organization of other task components (e.g., reward or stimuli). Thus, future studies can explore how brains and ANNs represent the abstraction of other task components. Second, though we were able to explore cross-context generalization across 64 contexts – significantly more than previous empirical studies in neuroscience – cross-context analysis was limited to a single task type (i.e., the C-PRO paradigm). It will be critical to see the organization of abstraction in multi-task settings that go beyond 64 contexts. Finally, our ANN modeling approach revealed the computational benefits of pretraining. It will be important for future work to benchmark sample efficiency and generalization performance against other training paradigms (e.g., in continual learning and/or meta-learning settings; [12, 44]).

In conclusion, we characterized a convergent hierarchical organization of abstract representations across the human cortex and in ANNs using a 64-context paradigm, and provided insight into the impact of abstract representations on generalization performance. Overall, we found that simple pretraining tasks efficiently embed abstract representations into ANNs, leading to improved systematic generalization similar to human behavior. These findings provide a human-centric benchmark from which to understand compositional generalization in ANNs, paving the way for greater interpretability of compositionality in ANNs. Importantly, investigating compositionality through a human-centric framework (e.g., by benchmarking ANNs against humans in the same task) creates a concrete target for interpreting the strengths and limitations of compositionality in ANNs. We hope these findings inspire further investigations into the comparison of compositionality in humans and ANNs.

# References


[1] Yashar Behzadi, Khaled Restom, Joy Liau, and Thomas T Liu. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, 37(1):90–101, 2007. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2007.04.042. URL http://www.sciencedirect.com/science/article/pii/S1053811907003837.

[2] Silvia Bernardi, Marcus K. Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C. Daniel Salzman. The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, October 2020. ISSN 00928674. doi: 10.1016/j.cell.2020.09.031. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867420312289.

[3] Rastko Ciric, Daniel H Wolf, Jonathan D Power, David R Roalf, Graham L Baum, Kosha Ruparel, Russell T Shinohara, Mark A Elliott, Simon B Eickhoff, Christos Davatzikos, Ruben C Gur, Raquel E Gur, Danielle S Bassett, and Theodore D Satterthwaite. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage*, 154:174–187, 2017. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2017.03.020. URL http://www.sciencedirect.com/science/article/pii/S1053811917302288.

[4] Michael W. Cole, Joset A. Etzel, Jeffrey M. Zacks, Walter Schneider, and Todd S. Braver. Rapid Transfer of Abstract Rules to Novel Contexts in Human Lateral Prefrontal Cortex. *Frontiers in Human Neuroscience*, 5:142, November 2011. ISSN 1662-5161. doi: 10.3389/fnhum.2011.00142. URL http://journal.frontiersin.org/article/10.3389/fnhum.2011.00142/abstract.

[5] Michael W. Cole, Patryk Laurent, and Andrea Stocco. Rapid instructed task learning: A new window into the human brain's unique capacity for flexible cognitive control. *Cognitive, Affective, & Behavioral Neuroscience*, pages 1–22, 2012. ISSN 1530-7026. doi: 10.3758/s13415-012-0125-7.

[6] Ronald Boris Dekker, Fabian Otto, and Christopher Summerfield. Determinants of human compositional generalization. Technical report, PsyArXiv, March 2022. URL https://psyarxiv.com/qnpw6/. type: article.

[7] Timo Flesch, Keno Juechems, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, 0(0), January 2022. ISSN 0896-6273. doi: 10.1016/j.neuron.2022.01.005. URL https://www.cell.com/neuron/abstract/S0896-6273(22)00005-8. Publisher: Elsevier.

[8] Steven M. Frankland and Joshua D. Greene. Concepts and Compositionality: In Search of the Brain's Language of Thought. *Annual Review of Psychology*, 71(1):273–303, January 2020. ISSN 0066-4308, 1545-2085. doi: 10.1146/annurev-psych-122216-011829. URL https://www.annualreviews.org/doi/10.1146/annurev-psych-122216-011829.

[9] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. J. Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2(4):189–210, 1994. ISSN 10659471. doi: 10.1002/hbm.460020402.

[10] Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, pages 1–11, 2016. ISSN 0028-0836. doi: 10.1038/nature18933. URL http://www.nature.com/doifinder/10.1038/nature18933.

[11] Matthew F Glasser, Stephen M Smith, Daniel S Marcus, Jesper L R Andersson, Edward J Auerbach, Timothy E J Behrens, Timothy S Coalson, Michael P Harms, Mark Jenkinson, Steen Moeller, Emma C Robinson, Stamatios N Sotiropoulos, Junqian Xu, Essa Yacoub, Kamil Ugurbil, and David C Van Essen. The Human Connectome Project's neuroimaging approach. *Nature neuroscience*, 19(9):1175–87, 2016. ISSN 1546-1726. doi: 10.1038/nn.4361. URL http://www.nature.com/neuro/journal/v19/n9/pdf/nn.4361.pdf%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/27571196.

[12] Raia Hadsell, Dushyant Rao, Andrei A. Rusu, and Razvan Pascanu. Embracing Change: Continual Learning in Deep Neural Networks. *Trends in Cognitive Sciences*, 24(12):1028–1040, December 2020. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2020.09.004. URL https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(20)30219-9. Publisher: Elsevier.

[13] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. November 2016. URL https://openreview.net/forum?id=Sy2fzU9gl.

[14] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a Definition of Disentangled Representations. *arXiv:1812.02230 [cs, stat]*, December 2018. URL http://arxiv.org/abs/1812.02230. arXiv: 1812.02230.

[15] Irina Higgins, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nat Commun*, 12(1):6456, November 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-26751-5. URL https://www.nature.com/articles/s41467-021-26751-5. Number: 1 Publisher: Nature Publishing Group.

[16] Drew A. Hudson and Christopher D. Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. *arXiv:1902.09506 [cs]*, May 2019. URL http://arxiv.org/abs/1902.09506. arXiv: 1902.09506.

[17] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality Decomposed: How do Neural Networks Generalise? *Journal of Artificial Intelligence Research*, 67:757–795, April 2020. ISSN 1076-9757. doi: 10.1613/jair.1.11674. URL https://www.jair.org/index.php/jair/article/view/11674.

[18] Takuya Ito and John D. Murray. Multi-task representations in human cortex transform along a sensory-to-motor hierarchy. Technical report, November 2021. URL https://www.biorxiv.org/content/10.1101/2021.11.29.470432v1. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

[19] Takuya Ito, Kaustubh R. Kulkarni, Douglas H. Schultz, Ravi D. Mill, Richard H. Chen, Levi I. Solomyak, and Michael W. Cole. Cognitive task information is transferred between brain regions via resting-state network topology. *Nat Commun*, 8, October 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-01000-w. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5715061/.

[20] Takuya Ito, Guangyu Robert Yang, Patryk Laurent, Douglas H. Schultz, and Michael W. Cole. Constructing neural network models from brain data reveals representational transformations linked to adaptive behavior. *Nat Commun*, 13(1):673, February 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-28323-7. URL https://www.nature.com/articles/s41467-022-28323-7. Number: 1 Publisher: Nature Publishing Group.

[21] Jie Lisa Ji, Marjolein Spronk, Kaustubh Kulkarni, Grega Repovš, Alan Anticevic, and Michael W Cole. Mapping the human brain's cortical-subcortical functional network organization. *NeuroImage*, 185: 35–57, 2019. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2018.10.006. URL http://www.sciencedirect.com/science/article/pii/S1053811918319657.

[22] W. Jeffrey Johnston and Stefano Fusi. Abstract representations emerge naturally in neural networks trained to perform multiple tasks. Technical report, October 2021. URL https://www.biorxiv.org/content/10.1101/2021.10.20.465187v2. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

[23] Daniel R. Kepple, Rainer Engelken, and Kanaka Rajan. Curriculum learning as a tool to uncover learning principles in the brain. September 2021. URL https://openreview.net/forum?id=TpJMvo0_pu-.

[24] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017. URL http://arxiv.org/abs/1412.6980. arXiv: 1412.6980.

[25] Brenden Lake and Marco Baroni. Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. In *International Conference on Machine Learning*, pages 2873–2882. PMLR, July 2018. URL http://proceedings.mlr.press/v80/lake18a.html. ISSN: 2640-3498.

[26] Brenden M. Lake. Compositional generalization through meta sequence-to-sequence learning. *arXiv:1906.05381 [cs]*, October 2019. URL http://arxiv.org/abs/1906.05381. arXiv: 1906.05381.

[27] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X16001837. URL https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/building-machines-that-learn-and-think-like-people/A9535B1D745A0377E16C590E14B94993. Publisher: Cambridge University Press.

[28] Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180, 2014.

[29] Daniel S. Margulies, Satrajit S. Ghosh, Alexandros Goulas, Marcel Falkiewicz, Julia M. Huntenburg, Georg Langs, Gleb Bezgin, Simon B. Eickhoff, F. Xavier Castellanos, Michael Petrides, Elizabeth Jefferies, and Jonathan Smallwood. Situating the default-mode network along a principal gradient of macroscale cortical organization. *PNAS*, 113(44):12574–12579, November 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1608282113. URL https://www.pnas.org/content/113/44/12574.

[30] Christian David Marton, Guillaume Lajoie, and Kanaka Rajan. Efficient and robust multi-task learning in the brain with modular task primitives. *arXiv:2105.14108 [cs, q-bio]*, May 2021. URL http://arxiv.org/abs/2105.14108. arXiv: 2105.14108.

[31] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

[32] Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. The role of Disentanglement in Generalisation. September 2020. URL https://openreview.net/forum?id=qbH974jKUVy.

[33] T E Nichols and Andrew P Holmes. Nonparametric Permutation Tests for Functional Neuroimaging Experiments: A Primer with examples. *Human Brain Mapping*, 15(1):1–25, 2001. ISSN 1065-9471. doi: 10.1002/hbm.1058. URL http://www3.interscience.wiley.com/cgi-bin/abstract/86010644/.

[34] Maxwell I. Nye, Armando Solar-Lezama, Joshua B. Tenenbaum, and Brenden M. Lake. Learning Compositional Rules via Neural Program Synthesis. *arXiv:2003.05562 [cs]*, March 2020. URL http://arxiv.org/abs/2003.05562. arXiv: 2003.05562.

[35] Steven T. Piantadosi. The Computational Origin of Representation. *Minds & Machines*, November 2020. ISSN 1572-8641. doi: 10.1007/s11023-020-09540-9. URL https://doi.org/10.1007/s11023-020-09540-9.

[36] Carlo Reverberi, Kai Görgen, and John-Dylan Haynes. Compositionality of Rule Representations in Human Prefrontal Cortex. *Cerebral Cortex*, 22(6):1237–1246, June 2012. ISSN 1460-2199, 1047-3211. doi: 10.1093/cercor/bhr200. URL https://academic.oup.com/cercor/article-lookup/doi/10.1093/cercor/bhr200.

[37] Jesse Rissman, Adam Gazzaley, and Mark D'Esposito. Measuring functional connectivity during distinct stages of a cognitive task. *NeuroImage*, 23(2):752–763, 2004. ISSN 10538119. doi: 10.1016/j.neuroimage.2004.06.035.

[38] Reidar Riveland and Alexandre Pouget. A neural model of task compositionality with natural language instructions. Technical report, bioRxiv, February 2022. URL https://www.biorxiv.org/content/10.1101/2022.02.22.481293v1. Section: New Results Type: article.

[39] Clemens Rosenbaum, Tim Klinger, and Matthew Riemer. Routing Networks: Adaptive Selection of Non-linear Functions for Multi-Task Learning. *arXiv:1711.01239 [cs]*, December 2017. URL http://arxiv.org/abs/1711.01239. arXiv: 1711.01239.

[40] Clemens Rosenbaum, Ignacio Cases, Matthew Riemer, and Tim Klinger. Routing Networks and the Challenges of Modular and Compositional Computation. *arXiv:1904.12774 [cs, stat]*, April 2019. URL http://arxiv.org/abs/1904.12774. arXiv: 1904.12774 version: 1.

[41] Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M Lake. A Benchmark for Systematic Generalization in Grounded Language Understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 19861–19872. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e5a90182cc81e12ab5e72d66e0b46fe3-Abstract.html.

[42] Luca Saglietti, Stefano Sarao Mannelli, and Andrew Saxe. An Analytical Theory of Curriculum Learning in Teacher-Student Networks. *arXiv:2106.08068 [cond-mat, stat]*, June 2021. URL http://arxiv.org/abs/2106.08068. arXiv: 2106.08068.

[43] Walter Schneider, Amy Eschman, and Anthony Zuccolotto. *E-Prime: User's guide*. Psychology Software Incorporated, 2002.

[44] Jane X. Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nat Neurosci*, 21(6):860–868, June 2018. ISSN 1546-1726. doi: 10.1038/s41593-018-0147-8. URL https://www.nature.com/articles/s41593-018-0147-8. Number: 6 Publisher: Nature Publishing Group.

[45] Guangyu Robert Yang, Igor Ganichev, Xiao-Jing Wang, Jonathon Shlens, and David Sussillo. A dataset and architecture for visual reasoning with a working memory. *arXiv preprint arXiv:1803.06092*, 2018.

[46] Guangyu Robert Yang, Madhura R. Joglekar, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, page 1, January 2019. ISSN 1546-1726. doi: 10.1038/s41593-018-0310-2. URL `https://www.nature.com/articles/s41593-018-0310-2`.

# Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] see Section 4.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] All human data has been de-identified and been publicly made available.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Public data URL is written in the text (Section 2.1)
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Applicable to Figures 1, 5, 6, 7
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Used previously published data [19]
   (b) Did you mention the license of the assets? [Yes] Published under a CC0 license, mentioned in section A.1.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Yes, see section A.1.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Yes, see section A.1. All data was previously de-identified.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] Yes – see Figure 1A
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] IRB approval was mentioned in section A.1
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No] This was a previously published dataset.

# A Supplementary Methods

## A.1 Participants

The following description is quoted with citation from a previous study that used the same dataset [20]. The dataset was publicly published under a CC0 license. All data on the public repository was previously de-identified.

Data were collected from 106 human participants across two different sessions (a behavioral and an imaging session). Technical error during MRI acquisition resulted in removing six participants from the study. Four additional participants were removed from the study because they did not complete the behavior-only session. fMRI analysis was performed on the remaining 96 participants (54 females). All participants gave informed consent according to the protocol approved by the Rutgers University Institutional Review Board. The average age of the participants that were included for analysis was 22.06, with a standard deviation of 3.84.

## A.2 C-PRO task paradigm – additional details

The C-PRO cognitive paradigm permutes specific task rules from three different rule domains (logical decision, sensory semantic, and motor response) to generate dozens of novel and unique task contexts. Visual stimuli included either horizontally or vertically oriented bars with either blue or red coloring. Simultaneously presented auditory stimuli included continuous (constant) or non-continuous (non-constant, i.e., "beeping") tones presented at high (3000Hz) or low (300Hz) frequencies. A given task context could be presented with 256 unique stimulus combinations. This is because a given task context was presented with two sequentially presented audiovisual stimuli, where each audiovisual stimulus varied in four dimensions: color (red/blue), orientation (vertical/horizontal), pitch (high/low), continuity (continuous/beeping). This led to $2^8 = 256$ possible stimulus combinations. The paradigm was presented using E-Prime software version 2.0.10.353 [43].

Each rule domain (logic, sensory, and motor) consisted of four specific rules, while each task context was a combination of one rule from each rule domain. A total of 64 unique task contexts (4 logic rules x 4 sensory rules x 4 motor rules) were possible, and each unique task set was presented twice for a total of 128 task miniblocks. This meant that there were $256 * 64 = 16384$ unique trials (i.e., context-stimulus) combinations. Identical task contexts were not presented in consecutive blocks. Each task miniblock included three trials, each consisting of two sequentially presented instances of simultaneous audiovisual stimuli. A task block began with a 3925 ms encoding screen (5 TRs), followed by a jittered delay ranging from 1570 ms to 6280 ms (2-8 TRs; randomly selected). Following the jittered delay, three trials were presented for 2355 ms (3 TRs), each with an inter-trial interval of 1570 ms (2 TRs). A second jittered delay followed the third trial, lasting 7850 ms to 12560 ms (10-16 TRs; randomly selected). A task block lasted a total of 28260 ms (36 TRs). Subjects were trained on four of the 64 task contexts for 30 minutes prior to the fMRI session. The four practiced rule sets were selected such that all 12 rules were equally practiced. There were 16 such groups of four task sets possible, and the task sets chosen to be practiced were counterbalanced across subjects. Subjects? mean performance across all trials performed in the scanner was 84% (median=86%) with a standard deviation of 9% (min=51%; max=96%). All subjects performed statistically above chance (25%).

## A.3 Analysis of human task performance data

The corresponding results described in this section can be found in Fig. 1.

We calculated the average accuracy for each task miniblock (comprising three task trials). Note that each task miniblock had the same task context (three-rule combination) for all three trials. This resulted in 128 task accuracy scores for every subject. Task rule contexts were then sorted into three categories separately for every subject ? practiced, 2-rule overlap, and 1-rule overlap. Practiced task contexts were defined as the four task contexts that were used to train participants on the C-PRO task outside of the MRI scanner. 2-rule overlap tasks were task contexts that had at least 2 of the same rules overlapping with the previously seen practiced tasks, and 1-rule overlap tasks were tasks with only a 1-rule overlap with practiced tasks. Note that there were no 0-rule overlap tasks, since subjects were trained on every rule prior to the test session. Moreover, there were no rule overlaps across practiced task contexts. Finally, every participant was provided with a randomly selected set of practiced task contexts. Behavioral accuracy was computed for every task context group for every subject (Fig. 1b).

We next evaluated the successive miniblock performance of each rule presented in a novel task context during the fMRI scanning session (see Fig. 1c-d). This would capture a participant?s ability to use a previously seen rule in a novel context as a function of the number of times it used the rule previously. The performance of an individual rule (e.g., "BOTH") was calculated separately per participant as a function of each novel context seen. Thus, performance of each rule was calculated for exactly 15 miniblocks (each rule was presented 16 times, including the practiced miniblock). For each rule, we then fit a linear regression model to assess whether performance (dependent variable) could be calculated as function of miniblock presentation (independent variable) with a

positive coefficient (i.e., increasing slope). A significant positive increase of performance (increasing positive performance) was tested for significance by assessing the p-value of the beta coefficient ($p<0.05$ threshold). This captured compositional learning ? re-using a task rule (despite use in a novel context) indicated that participants were learning to use previously learned rules in out-of-set novel contexts.

### A.4 fMRI acquisition

The following fMRI acquisition details are taken from a previous study that used the same data set [20].

Whole-brain multiband echo-planar imaging (EPI) acquisitions were collected with a 32-channel head coil on a 3T Siemens Trio MRI scanner with TR=785 ms, TE=34.8 ms, flip angle=55°, Bandwidth 1924/Hz/Px, in-plane FoV read=208 mm, 72 slices, 2.0 mm isotropic voxels, with a multiband acceleration factor of 8. Whole-brain high-resolution T1-weighted and T2-weighted anatomical scans were also collected with 0.8 mm isotropic voxels. Spin echo field maps were collected in both the anterior to posterior direction and the posterior to anterior direction in accordance with the Human Connectome Project preprocessing pipeline [11]. A resting-state scan was collected for 14 minutes (1070 TRs), prior to the task scans. Eight task scans were subsequently collected, each spanning 7 minutes and 36 seconds (581 TRs). Each of the eight task runs (in addition to all other MRI data) were collected consecutively with short breaks in between (subjects did not leave the scanner).

### A.5 fMRI preprocessing

The following details are quoted with citation from a previous study that used the same preprocessing scheme [20].

Resting-state and task-state fMRI data were minimally preprocessed using the publicly available Human Connectome Project minimal preprocessing pipeline version 3.5.0. This pipeline included anatomical reconstruction and segmentation, EPI reconstruction, segmentation, spatial normalization to standard template, intensity normalization, and motion correction64. After minimal preprocessing, additional custom preprocessing was conducted on CIFTI 64k grayordinate standard space for vertex-wise analyses using a surface based atlas [11]. This included removal of the first five frames of each run, de-meaning and de-trending the time series, and performing nuisance regression on the minimally preprocessed data [3]. We removed motion parameters and physiological noise during nuisance regression. This included six motion parameters, their derivatives, and the quadratics of those parameters (24 motion regressors in total). We applied aCompCor on the physiological time series extracted from the white matter and ventricle voxels (5 components each extracted volumetrically) [1]. We additionally included the derivatives of each component time series, and the quadratics of the original and derivative time series (40 physiological noise regressors in total). This combination of motion and physiological noise regressors totaled 64 nuisance parameters, and is a variant of previously benchmarked nuisance regression models [3].

### A.6 fMRI activation estimation

We performed a within-subject task GLM on the vertex-wise fMRI time series to estimate task rule-related activations on the CIFTI grayordinate space. To extract task activations for each task block, we performed a beta series regression on every task miniblock [37]. Specifically, we fit an independent regressor to every encoding period (3925ms, 5 TRs), resulting in 128 task regressors in total. Fitting regressors on the encoding period was done primarily to isolate rule representations rather than the actual trial (stimulus-response) period. Each regressor was a boxcar function that was a vector of 0s every except for the specified encoding period. This boxcar function was then convolved with the SPM canonical hemodynamic response function [9]. A single activation estimate (beta coefficient) was extracted for every encoding block at every surface vertex.

### A.7 ANN construction and batch training

The primary ANN architecture was comprised of two hidden layers, each with 128 units. The output layer was comprised of four units that corresponded to each motor response. The ANN transformed the trial input vector into a 4-element response vector with the equation $Y = f_{ReLU}(X_h W_h + b_h)$, where $Y$ corresponds to the output vector, $W_h$ is the weight matrix from the last hidden layer to the output layer, $X_h$ is the activation vector of the last hidden layer, and $b_h$ is the bias vector. The hidden unit activation vectors were defined as $X_i = f_{ReLU}(X_{i-1} W_{i-1} + b_{i-1})$, where $X_i$ is the activation vector for layer $i$. Weights and biases were initialized from a uniform distribution $U(-\sqrt{1/k}, \sqrt{1/k})$, where $k$ is the number of input features from the previous layer.

The ANN was optimized by minimizing the cross entropy between the outputs and the correct target output (a one-hot vector). Optimization was performed using Adam with a learning rate of 0.001 [24]. During training, we used dropout (with probability 0.2 from input to hidden, and 0.5 within hidden layers).

Training on the C-PRO task was performed in a sequential learning paradigm. Initially, an arbitrary set of four practiced contexts were selected with the constraint that no set of four practiced contexts had any overlapping

rules. (This was randomly selected across different ANN initializations). Then, novel task contexts were incrementally added (by 1) into the set of training contexts, and performance (and ANN analysis) was performed after the addition of each task context. Training stopped once all 64 task contexts had been fully trained on. We used batch training. Each batch contained a single task context with all possible stimulus (256) combinations. Thus, each batch contained 256 trials in total.

To stop training we set criteria for two different experiments. The first experiment required that performance on each task context (each batch) in the training set achieved a baseline performance accuracy or better (90%). Once this criterion was satisfied, a novel task context would be added into the training. The second experiment kept fix the number of batches/gradient steps each task context took prior to adding a new task context. This was set to 200 gradient steps per task context.

### A.8 ANN pretraining description

There were two pretraining paradigms: Primitives pretraining and Simple task pretraining.

Primitives pretraining focused on training individual rules within each rule domain, enabling ANNs learn individual rule Primitives (e.g., the notion of what "RED" is), and is consistent to how humans enter the full C-PRO experiment – most participants already know what the primitives "RED", "LEFT MIDDLE", or "BOTH" refer to. While C-PRO task contexts activated 3 rules out of 12 possible task rules, Primitives tasks only activated 1/12 rules. Primitives pretraining was performed for each of the rule domains separately.

Motor Primitives pretraining involved making motor responses given a motor rule. This involved activating one motor rule in the input vector at a time. If the "LEFT MIDDLE" rule was activated, a LEFT MIDDLE output response would be expected. In addition, if the "NOT" unit was activated in conjunction with the "LEFT MIDDLE" rule, then the "LEFT INDEX" output response would be expected, which is the analogous rule instruction that participants received prior to performing the C-PRO task.

Sensory Primitives pretraining involved making True/False statements on whether a specific sensory stimulus feature was presented. This involved activating one sensory rule and one sensory stimulus feature at a time. For example, for the sensory rule "RED", either a RED/BLUE sensory stimulus would activate. During pretraining, we included two additional output units that corresponded to True/False units. If a "RED" stimulus was presented in conjunction with "RED" sensory rule, the output should be "True"; otherwise, "False". In addition, we also presented sensory rules with the "NOT" negation. In other words, if the rule was "NOT RED" and the sensory stimulus presented was RED, then the network should produce "False".

Logic Primitives pretraining involved learning the abstract logical relations between different logic rules. The logic rules "BOTH" and "EITHER" could be equivalently interpreted as "AND" and "OR" logic rules, respectively. The "NOT BOTH" and "NEITHER" rules were analogous to the negations of those rules. This was operationalized by presenting two stimuli from a specific feature domain (e.g., color). Using stimuli from the color feature domain as an example, for "BOTH" Primitive training, if RED-then-RED stimuli were presented, this would result in a "True" boolean. In contrast, if RED-then-BLUE stimuli were presented, then this would result in a "False" boolean. This intuition was derived from conditional logic (where "BOTH" is equivalent to the concept of "SAME"), where the statement $x == x$ is True. In contrast, "EITHER" (i.e., "OR") was coded as "True" for any combination features of a color stimuli. The negations "BOTH" and "EITHER" resulted in the negation of the produced boolean (i.e., "NEITHER" produced "False" for all stimulus pairs). The most critical point, however, is that the responses for all logic rules were distinct among each other, and described equivalent logical relations between rules. Previous theoretical work in cognitive science suggests that symbolic computation emerges by learning the relational representations between symbolic operations [35]. This logic Primitives pretraining approach therefore captures the relations amongst symbolic operations.

Simple task pretraining involved combining two-rule task context, rather than the full three-rule task context in the C-PRO task. We designed two variants of simple tasks: a sensorimotor task that combined a sensory and motor rule, and a logical-sensory task that combined a logic and sensory rule. The sensorimotor task activated a sensory rule, motor rule, and one sensory stimulus unit. The response output was a motor response. For example, if the rules were "RED" and "LEFT MIDDLE", and the stimulus was "RED", ANNs were taught to respond with the Left Middle response unit. If the stimulus was anything other than "RED", then ANNs were taught to respond with the Left Index response unit. The sensorimotor task was used to teach ANNs simple mappings between sensory input and motor responses in the simplest possible paradigm.

The logical-sensory task was designed to teach ANNs logical inferences over sensory stimuli in the simplest possible manner. Like the sensorimotor task, the logical-sensory task included two rules: a logic and sensory rule. However, unlike the sensorimotor task, it activated two stimulus units from the same feature domain (e.g., the color domain). For example, for the logic rule was "BOTH" and sensory rule "RED", if the first stimulus was red and second stimulus red, the ANN would be taught to respond with the True output unit. If any of the stimuli were not red (e.g., blue), then the ANN would be taught to respond with the False output unit.

Primitives pretraining was always performed prior to Simple task pretraining, except for Fig. 12, which investigated the effect of reversing the order of pretraining tasks. Pretraining procedures were blocked together, such that all conditions within the Primitives pretraining paradigm (i.e., Logic, Sensory, and Motor primitives pretraining) were trained until all three tasks achieved 99% accuracy. Simple task pretraining was subsequently performed until both Logical-Sensory and Sensory-Motor tasks were performed at 99% accuracy. This ordering is consistent with prior work, suggesting that ANNs are more sample efficient when transitioning from easier to more difficult tasks [42]. We also performed a simple control experiment demonstrating that when pretraining was reversed in the Combined condition (i.e., Simple task pretraining followed by Primitives pretraining), generalization performance was reduced to chance (Fig. 12). This suggests that the ordering of pretraining paradigms is crucial for generalization performance and sample efficiency, which future work should explore.

After pretraining, the additional True/False output units were lesioned from the network.

### A.9 ANN analysis

Analysis of ANNs was carried out in a similar manner to how empirical fMRI data and behavior was analyzed. ANN analysis was performed to infer how the structure of their internal representations was associated with task generalization performance and sample efficiency. Task generalization performance was calculated as the performance on novel contexts (i.e., novel recombinations of task rules). This was independent of whether or not ANNs had learned/seen individual rules previously. To estimate task sample efficiency, we calculated the number of trial samples required to achieve a baseline task accuracy percentage (on contexts in the training set). Note that for a fixed number of trial samples, the number of gradient steps were the same (batch sizes were always the same). Generalization inefficiency was measured as the ratio of the number of samples trained on (normalized between 0 and 1) and the generalization performance on novel contexts (normalized to 0 to 1 + a fixed constant).

PS in ANNs was calculated separately from the training procedure. Since we were only interested in the PS of rule representations, only input units associated with the task rules were activated, while stimulus inputs were set to 0. This ensured that the hidden activations were not contaminated by stimulus-related activations when calculating PS in the hidden layers. Otherwise, PS in ANNs was calculated in a similar manner to the empirical fMRI data, where the spatial features (dimensions) were the units within a given hidden layer (like voxels within a brain parcel).

In addition, Figures 15, 16, and 17 illustrate that pairwise PS scores for all pairs of rule dichotomies in both ANNs and human fMRI data.

### A.10 Computing resources

fMRI analyses were carried out on a local server with 24 cores and 320GB of RAM. ANN training, while not required, was performed on an Nvidia P100 GPU. A single ANN initialization can be successfully trained on a CPU in under 2 minutes.

## B Supplementary Figures

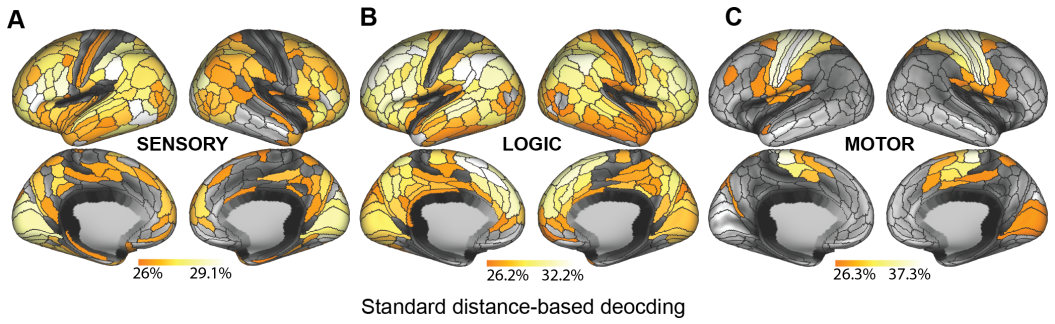


Figure 7: Standard decoding of multivariate activation patterns reveals more distributed patterns of decodability than identified with PS. These findings suggest that of the regions that contain task information, only a subset of these regions contain abstract representations. a) Sensory, b) logic, and c) motor rule decoding at the group level (n=96). Decoding was performed using a distance-based classifier (pearson similarity), and significance was assessed using a binomial test against chance (25%). Significance was assessed using multiple comparisons corrected threshold (False Discovery Rate) of $p<0.05$.

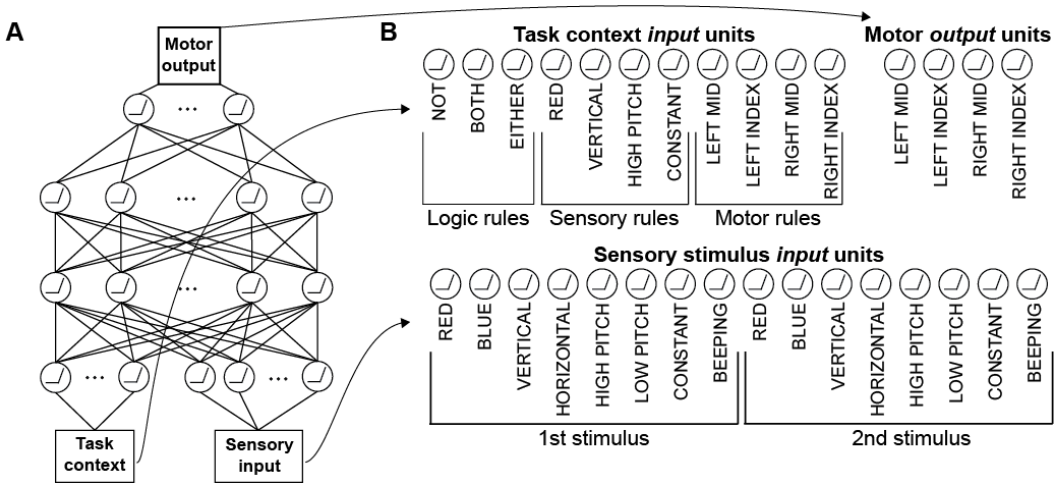


Figure 8: a) ANN architecture. We employed a simple architecture with two hidden layers (128 units each) with rectified linear units. b) The ANN?s input and output units. Task input units were modeled as one-hot vectors for each rule across all rule domains. Sensory stimulus input units were modeled as categorical one-hot vectors for each stimulus type. Separate vectors were modeled for first and second stimulus presentations. Motor outputs had a unit corresponding to each motor finger response.

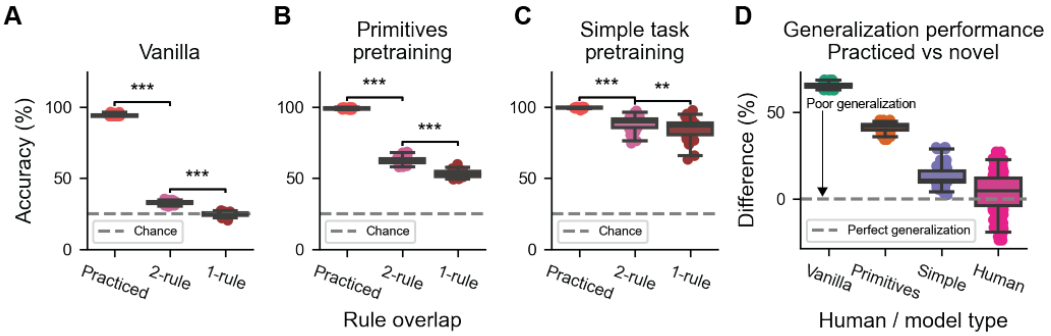


Figure 9: a-c) We trained the ANN architecture on 4/64 C-PRO task contexts with (a) a vanilla ANN with randomly initialized weights, (b) an ANN after Primitives pretraining, and (c) an ANN after Simple task pretraining. Training on the 4 practiced contexts was stopped after the model achieved 90% accuracy on those contexts. As in our empirical human data (Fig. 1b), the ANN performed best on contexts in which it had previously seen (Practiced), followed by contexts with two overlapping (2-rule) and one overlapping (1-rule) rules. d) Generalization performance of ANNs with Primitives and Simple task was most similar to human generalization performance in novel contexts. (Lower values indicate better generalization.)

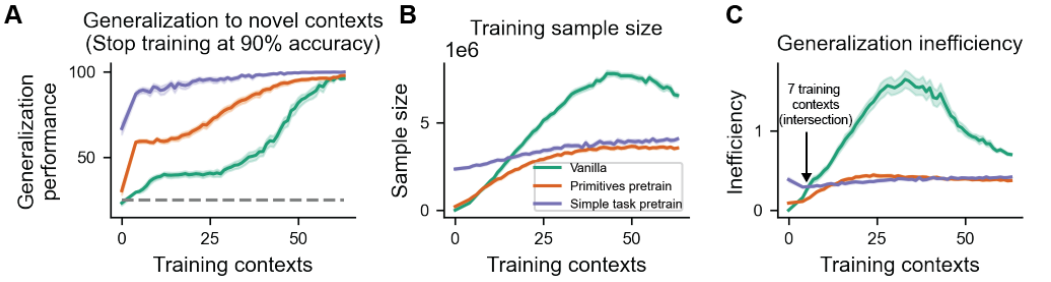


Figure 10: a) Generalization performance on novel contexts after training on n/64 contexts (x-axis). b) Number of total trials/samples shown to each model type. c) The generalization inefficiency of each model (the normalized ratio of the number of samples shown and the performance on novel tasks). See Fig. 11 for model performance using a fixed number of samples per trained context.

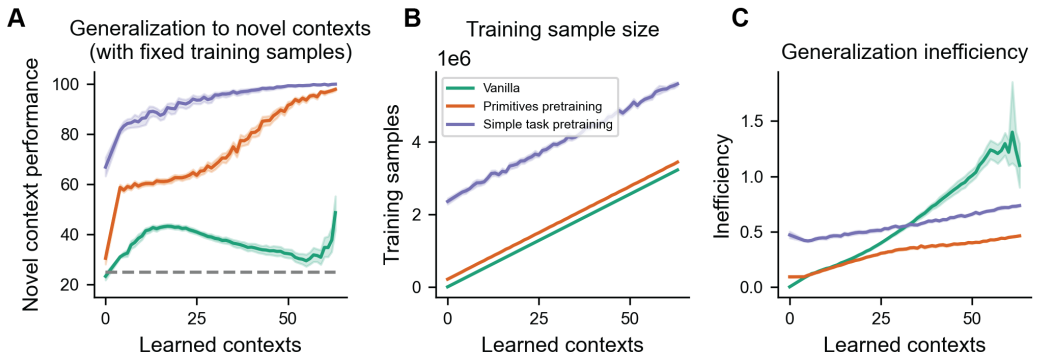


Figure 11: Novel context performance as function of learned contexts in the C-PRO task in the ANN model. Similar to Fig. 10, but we trained the model using a fixed number of samples per learned context (200 epochs per context) and evaluated generalization performance. a) We systematically trained each type of ANN on a range from 0-63 of the C-PRO task contexts, and assessed its generalization on the remaining (novel) contexts. We trained the model on a fixed number of samples per learned context. Learned contexts were sequentially introduced, and novel context performance was assessed on the remaining (excluded from training) contexts. b) The number of total trials/samples shown to each model type. Since the number of samples shown to the model for each context was fixed, training samples linearly increased as a function of learned contexts. c) As in Fig. 10c, we measured the generalization inefficiency of each model. Generalization inefficiency was measured as the ratio of the number of samples shown (normalized between 0 and 1) and the performance on novel tasks (normalized to 0 to 1 + a fixed constant). Thus, while the vanilla model was initially more efficient, pretrained models quickly learned more efficiently as evidenced by better generalization accuracies with fewer samples.

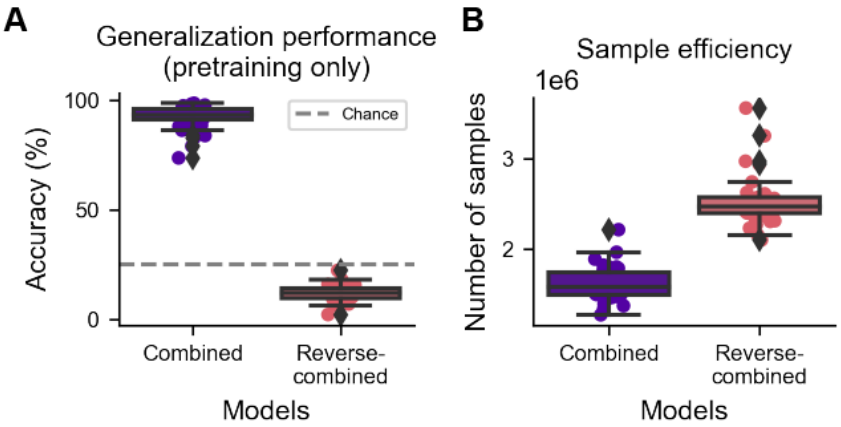


Figure 12: Order of pretraining paradigms drastically affects generalization performance and sample efficiency. In the Combined pretraining regimen, we first implemented Primitives pretraining followed by Simple task pretraining. However, prior curriculum research suggests that the order of task learning can significantly impact generalization performance [42]. Thus, we compared the Combined pretraining protocol to a Reverse-combined protocol, where Simple task pretraining was performed prior to Primitives pretraining. a) We found that generalization performance on the unseen C-PRO tasks was significantly impaired in the Reverse-combined condition. b) We also measured the sample efficiency of only the pretraining trials. We found that Reverse-combined pretraining required more samples than Combined pretraining, despite the stopping criterion for each pretraining task remaining the same. (ANNs were required to perform either Primitives pretraining or Simple task pretraining with at least 99% accuracy prior to moving on to the next pretraining task.) This suggests that ordering of pretraining tasks can significantly impact the learning and generalization dynamics of simple ANNs.

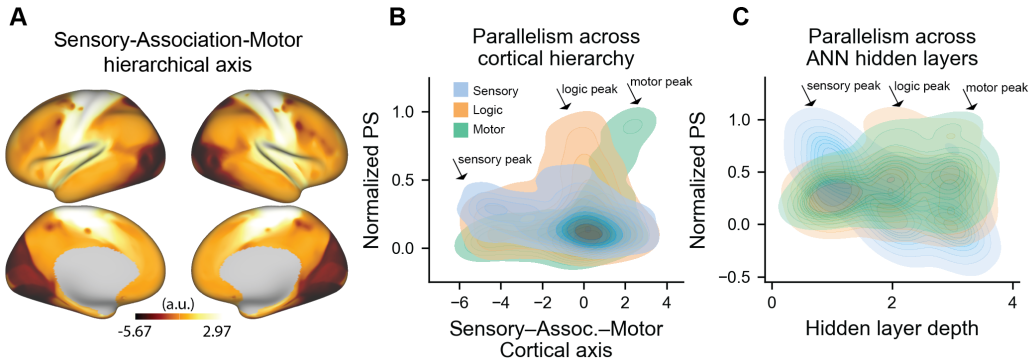


Figure 13: Convergent hierarchical organization of abstract representations in humans and ANNs. a) We compared the topographic differences in PS across cortex with a well-known sensory-to-motor hierarchical gradient identified during resting-state fMRI [29]. b) We compared how content-specific abstractions (PS) differed across this hierarchical gradient, finding that sensory rule abstractions were highest in the lower part of the gradient (sensory systems), logic rule abstractions were highest across association cortex, and motor rule abstractions were highest across motor cortex. c) For an analogous analysis, we plotted how domain-specific abstraction differed across different hidden layer depths in the pretrained ANN with 3 hidden layers. We identified similar patterns of parallelism across the 3 hidden layers for each of the rule domains.

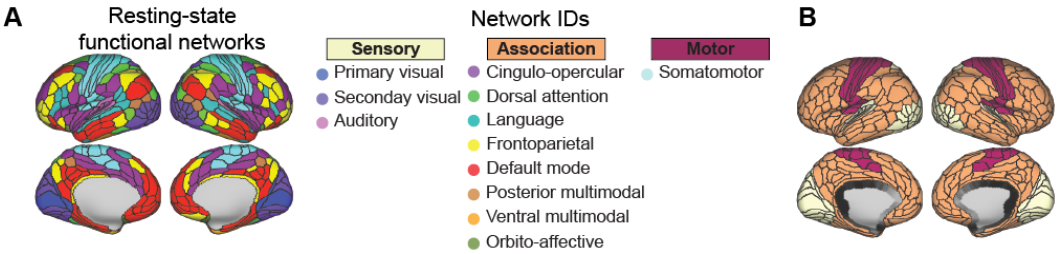


Figure 14: a) A network partition of cortical parcels using resting-state fMRI [21]. b) To match the number of layers in the ANN, we created three discretized systems based on the functional network partition – sensory, association, and motor – that followed the sensory-to-motor hierarchy (Fig. 6) [29].

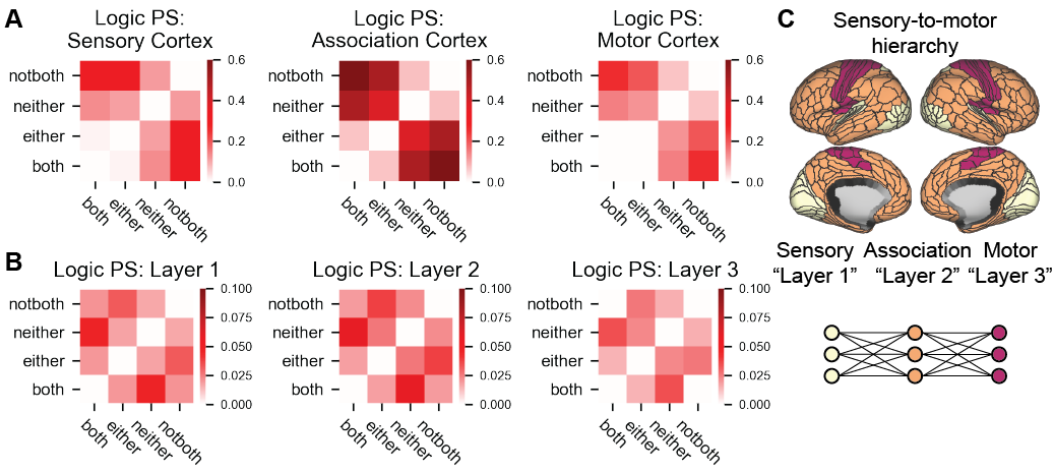


Figure 15: A pairwise comparison of PS for all possible dichotomies in the Logic rule condition. a) We computed the PS for every pairwise dichotomy in the Logic rule domain, and computed the average PS across brain regions within each cortical system (Sensory, Association, and Motor cortex). The average PS for each cortical system was computed by averaging the PS across all regions within that cortical system for every dichotomy. b) To compare how the pairwise dichotomies matched in pretrained ANNs (Combined pretrained only; no training on full C-PRO contexts), we computed the PS for all dichotomies in each layer. In this particular experiment we used an ANN with three hidden layers to compare with Sensory, Association, and Motor cortical systems. c) The sensory, association, and motor cortical systems can be analogized to hidden layer depths in the ANN.

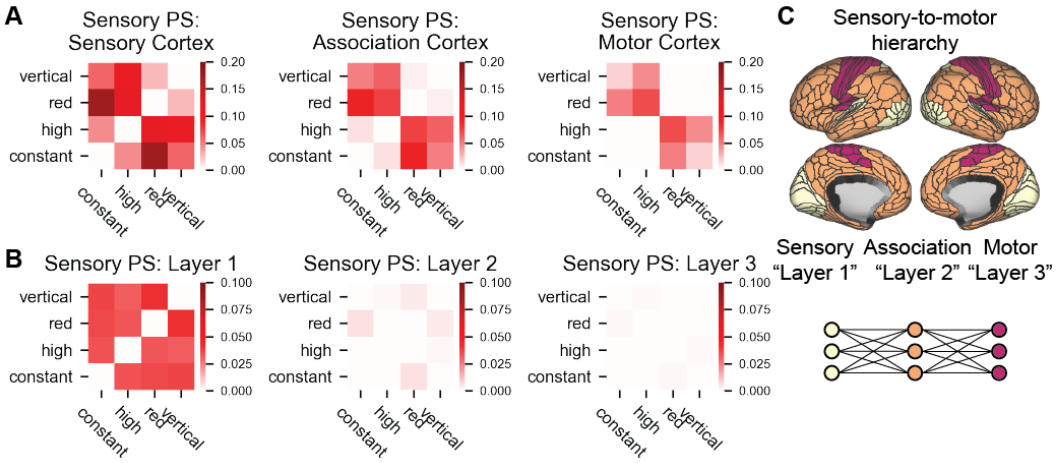


Figure 16: A pairwise comparison of PS for all possible dichotomies in the Sensory rule condition. a) We computed the PS for every pairwise dichotomy in the Sensory rule domain, and computed the average PS across brain regions within each cortical system (Sensory, Association, and Motor cortex). The average PS for each cortical system was computed by averaging the PS across all regions within that cortical system for every dichotomy. b) To compare how the pairwise dichotomies matched in pretrained ANNs (Combined pretrained only; no training on full C-PRO contexts), we computed the PS for all dichotomies in each layer. In this particular experiment we used an ANN with three hidden layers to compare with Sensory, Association, and Motor cortical systems. c) The sensory, association, and motor cortical systems can be analogized to hidden layer depths in the ANN.

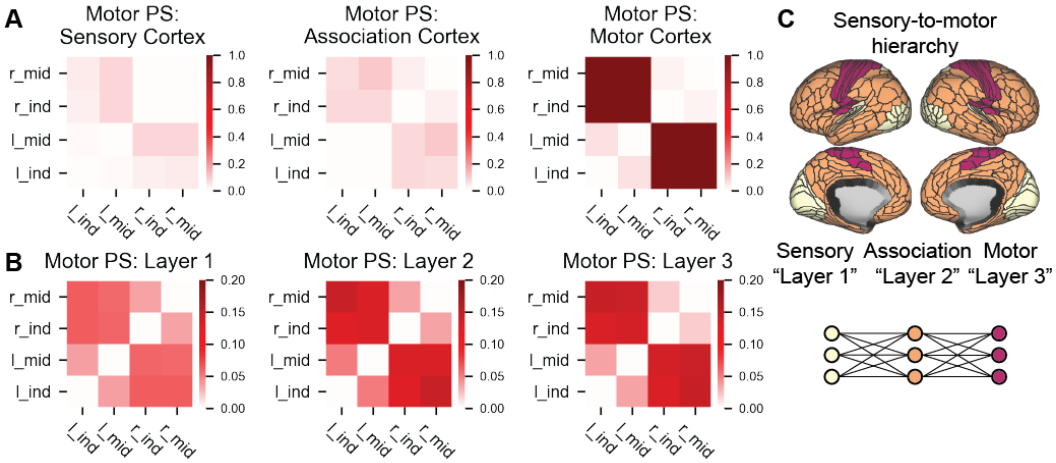


Figure 17: A pairwise comparison of PS for all possible dichotomies in the Motor rule condition. a) We computed the PS for every pairwise dichotomy in the Motor rule domain, and computed the average PS across brain regions within each cortical system (Sensory, Association, and Motor cortex). The average PS for each cortical system was computed by averaging the PS across all regions within that cortical system for every dichotomy. b) To compare how the pairwise dichotomies matched in pretrained ANNs (Combined pretrained only; no training on full C-PRO contexts), we computed the PS for all dichotomies in each layer. In this particular experiment we used an ANN with three hidden layers to compare with Sensory, Association, and Motor cortical systems. c) The sensory, association, and motor cortical systems can be analogized to hidden layer depths in the ANN.